**Sound**

This chapter introduces another basic perceptual sensation: Sound. Hearing and vision are the two most important sensual inputs that humans can process. Many parallels exist between visual signal processing and acoustic signal processing, but sound has unique properties—often **orthogonal** to those of visual signals. Perhaps this is why nature gave animals both visual and acoustic sensors: To gather complementary information about the environment. Many species use sound to detect danger, navigate, locate prey, and communicate. Earth's atmosphere and water, as well as virtually all physical phenomena—fire, rain, wind, surf, earthquake, and so on—produce unique sounds. Species of all kinds, such as frogs, birds, marine and terrestrial mammals, have developed special organs to produce sound. In some species, these have evolved to produce singing and speech. Furthermore, humans have developed culture and technology (for example music, telephone, and radio) that let them generate, record, transmit, and broadcast sound.

In this chapter, we introduce the basic properties of sound, sound production, and sound perception. Of course, a multimedia book can only scratch the surface of this complex and fascinating topic.

**[A head] What is sound?**

The *American Heritage Dictionary of the English Language*, Fourth Edition defines sound as “a traveling wave which is an oscillation of pressure transmitted through a solid, liquid, or gas, composed of frequencies within the range of hearing and of a level sufficiently strong to be heard, or the sensation stimulated in organs of hearing by such vibrations.” This compressed formulation is perfect start for discussing the properties of sound. Sound is generated by mechanical oscillation. Unlike light, sound must travel through a medium. In a vacuum there is no sound so, for example, one can’t hear exploding space ships. The traveling speed of sound varies according to the medium it is traveling in. In dry air at 20 °C, the speed of sound is 343 meters per second, or Mach 1. In fresh water, also at 20 °C, the speed of sound is approximately 1,482 meters per second.

For a sound to be heard, the oscillation’s frequency and amplitude must be in a certain range. For humans, hearing is normally limited to frequencies between 12 and 20,000 Hz (20 kHz). The upper limit generally decreases as the person ages. Other species have a different range of hearing. Dogs, for example, can perceive vibrations higher than 20 kHz. This is also one reason why dogs and cats will not react to broadcast TV the same way humans do. Even though **//te squeak of?//** a mouse can draw a cat’s attention from hundreds of meters away, the same sound, much more intense, from an MP3 player or a TV might not interest the cat at all.

Because sound is an oscillation of pressure, you can measure a sound wave’s amplitude by measuring the *sound pressure*. Sound pressure is defined as the difference between the average local pressure of the medium outside of the sound wave it is traveling through (at a given point and a given time) and its pressure within the sound wave **//correct?//**. The square of this difference is usually averaged over time and/or space. By taking a square root of the average, you obtain a root mean square (RMS) value.

The sound pressure perceived by the human ear is nonlinear (see Chapter XXX [mulaw]), and the range of amplitudes is rather wide. Therefore, sound pressure is often measured as a level on a logarithmic scale, the *decibel scale*. The sound pressure level (SPL) or *L*p is defined as:



where *p* is the RMS sound pressure and *p*ref is a reference sound pressure. Commonly used reference sound pressures for silence, defined in the standard ANSI S1.1-1994, are 20 µPa in air and 1 µPa in water. Most sound recording equipment is calibrated to omit 0-amplitude at these levels.

The human ear does not have a flat spectral response (we discuss this in more detail later in the chapter). That is, the same sound pressure at a different frequency will be perceived as a different volume level. Therefore, sound pressures are often frequency weighted so the measured level will more closely match perceived levels. The *A-weighting scheme*, defined by International Electrotechnical Commission (IEC) and illustrated in Figure 1, is the most common.

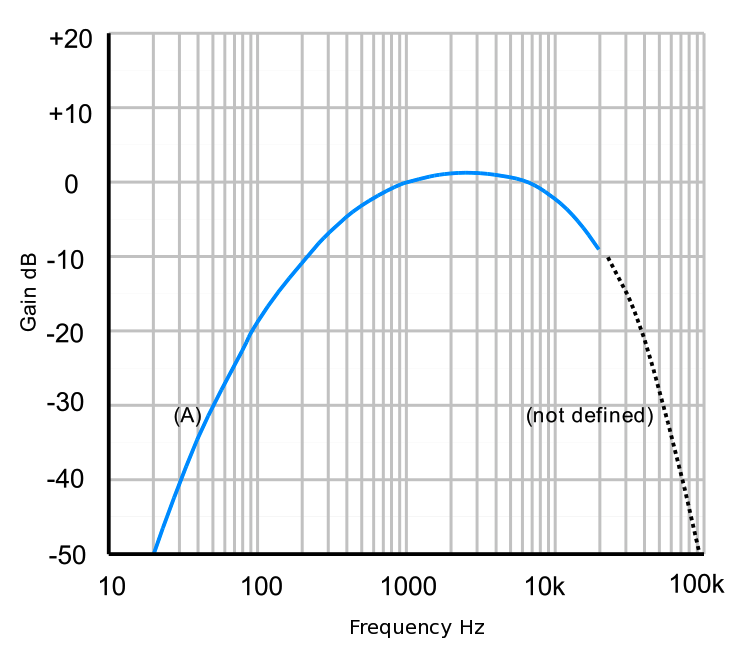


Figure 1. Sound pressure A-weighting scheme according to IEC 61672:2003.

Sound pressure levels weighted by this scheme are usually labeled as dBA or dB(A). The terms dB and dBA, like the percent symbol (%), define ratios rather than physical measurement units. A value of 10 dB can refer to completely different sound pressure levels, depending on the reference.

**[A Head] Observed Properties of Sound**

In practice, sounds do not travel exclusively in a homogenous medium from a source to exhaustion. The environment is full of objects that sound can travel through or be reflected from. Sound pressure waves can collide with each other. The resulting effects of these conditions play a large role in the design of multimedia systems. In addition, the effects **on //of? or the effects of sound on what?//** sound are more significant than **//what is?//** on light waves. The three most important sound effects are echo, reverberation, and interference.

An *echo* is a reflection of sound, arriving at the listener some time after the original sound. Typical examples are the echo produced by the bottom of a well, by a building, or by the walls of a closed-in room. Because most materials easily reflect sounds, echoes are always present in every environment. The time delay is the extra distance divided by the speed of sound. The human ear cannot distinguish an echo from the original sound if the delay is less than 1/10 of a second. Thus, because the velocity of sound is approximately 343 meters per second at a normal room temperature of about 20°C, the reflecting object must be more than 16.2 meters from the sound source at this temperature for a person at the source to hear an echo. For a sound wave to travel that far back and forth it must have sufficient energy. Normal conversation is usually below this energy threshold.

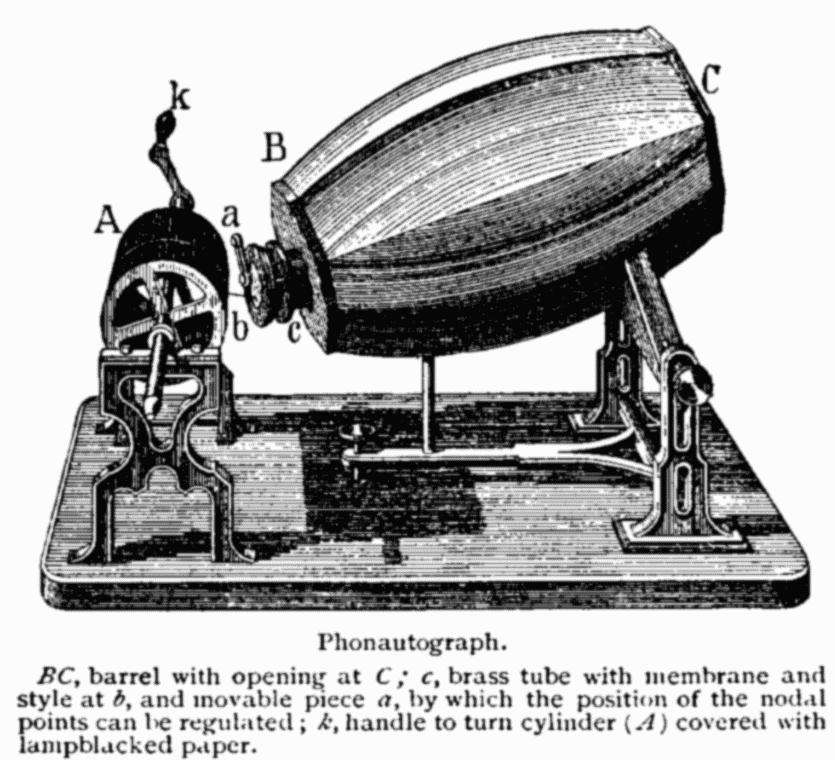


Figure 2. Phonautograph by Édouard-Léon Scott de Martinville. (Source: Uncredited 19th century engraving, Wikimedia Commons)

*Reverberation* is the persistence of sound in a particular space after the original sound wave is exhausted. A reverberation is created when a sound is produced in an enclosed space, causing numerous echoes to build up and then slowly decay as the environment absorbs the sound. This is most noticeable to the human ear when the sound source stops but the reflections continue, decreasing in amplitude, until they are no longer audible. Reverberation receives special consideration in the architectural design of large chambers, which need specific reverberation times to achieve optimum performance for their intended activity. Unless a room and recording equipment is specially designed to not cause reverberation, reverberation is always present. Reverberation is also present during the production of speech in the vocal tract. Reverberation is characterized by the *reverberation time*, that is, the length of this sound decay. Multimedia content analysis techniques often suffer from not accounting for reverberation, even when it is inaudible.

*Interference* is the superposition of two or more waves that results in a new wave pattern. It usually refers to the interaction of waves that are correlated or coherent with each other, either because they have the same source or the same or nearly the same frequency. Sound interference causes different effects, which are described in wave propagation equations in physics.

Multimedia system designers should be aware of interference, which they can use constructively and destructively.

Consider two waves that are in phase, with amplitudes A1 and A2. Their troughs and peaks line up and the resultant wave will have amplitude A = A1 + A2. This is known as *constructive interference*.

If the two waves are 180° out of phase, one wave's crests will coincide with another wave's troughs and so will tend to cancel **//each other///** out. The resultant amplitude is A = |A1 − A2|. If A1 = A2, the resultant amplitude will be zero. This is known as *destructive interference*. **//designers?//** often use destructive interference to eliminate unwanted sounds—for example, in noise-canceling earphones. However, this is not the same as masking effects, which are caused by the perceptual properties of the human brain (and will be discussed in chapter XX).

**[A Head] Recording and Reproduction of Sound**

Humans have tried to accurately record sound for thousands of years **//???//**. The first device that could record sound mechanically (but could not play it back) was the phonautograph, developed in 1857 by Parisian inventor Édouard-Léon Scott de Martinville. This machine produced *phonautograms*, the earliest known recordings of the human voice **//correct?//**. These earliest known recordings include a dramatic reading in French of Shakespeare’s Othello and music played on a guitar and trumpet.

Figure 2 shows a schematic of the device from the inventor’s original records. A barrel with an opening captured the sound waves and focused them onto a membrane attached to a hog's bristle **//correct?//**, causing the bristle to move and allowing it to inscribe the sound onto a visual medium. Even though this device was more an early oscillograph than a sound recording device, the concept of the first practical sound recording and reproduction device wasn’t too different.

Thomas A. Edison invented the mechanical phonograph cylinder in 1877 and patented it in 1878. The recordings were initially stored on the outside surface of a strip of tinfoil wrapped around a rotating metal cylinder. To play back the recordings, a needle ran along the cylinder, applying less pressure than in the recording to convert the mechanical engravings into sound waves that would be mechanically amplified **//rewording correct?//**. Figure 3 shows a US postage stamp featuring the device.



Figure 3. Edison’s phonograph on a US postage stamp.

Not surprisingly, sound recording still obeys the same principles, with two main exceptions.

* The sound waves are converted to electrical waves by a microphone, and
* Most of today’s storage media is digital—that is, sound waves are converted to binary numbers before they are imprinted on the medium.

The media themselves, such as CDROM or DAT, are a bit more sophisticated than Edison’s cylinders. Currently, generic media, such as hard disks and flash memory, are replacing these specialized media.

The next sections explain the governing principles of modern sound processing.

**[B Head] Microphones**

A microphone is an acoustic sensor that converts sound into an electrical signal. The general principle is that sound pressure is inflicted on a membrane that varies its electrical resistance according to the movement. Most current microphones **//Okay to cut “in use for audio”? What else would they be used for?//** use electromagnetic induction (dynamic microphone) by letting the membrane swing a magnetic field produced by a coil, capacitance change (condenser microphone) by letting the membrane be part of a capacitor that varies capacity with movement, or piezoelectric generation (piezo crystals emit electricity when under pressure). Modern laser microphones use light modulation to produce the electric signal by “watching” the mechanical vibration.

A single dynamic membrane will not respond linearly to all audio frequencies. For this reason some microphones use multiple membranes for the different parts of the audio spectrum and then combine the resulting signals. The different microphone types have different electrical properties. A complete microphone also includes a housing and a means of bringing the signal from the element to other equipment (such as wires or RF capability). These and other characteristics of the microphone, such as diaphragm size, intended use, or orientation of the principal sound input to the principal axis (end- or side-address), determine the properties of the recorded sound space. This, when planning a recording, it is best to survey what is available in the market and read vendor specifications.

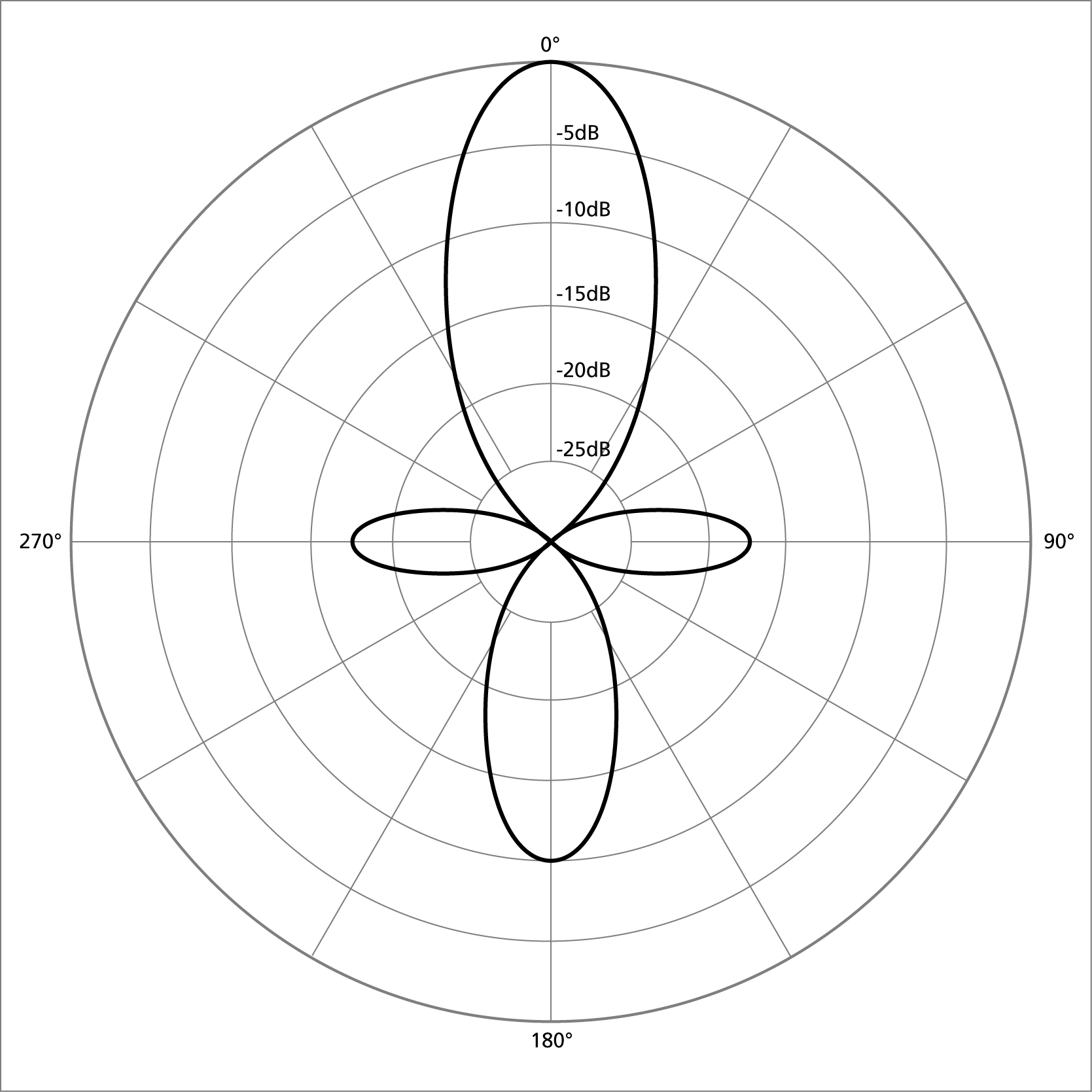
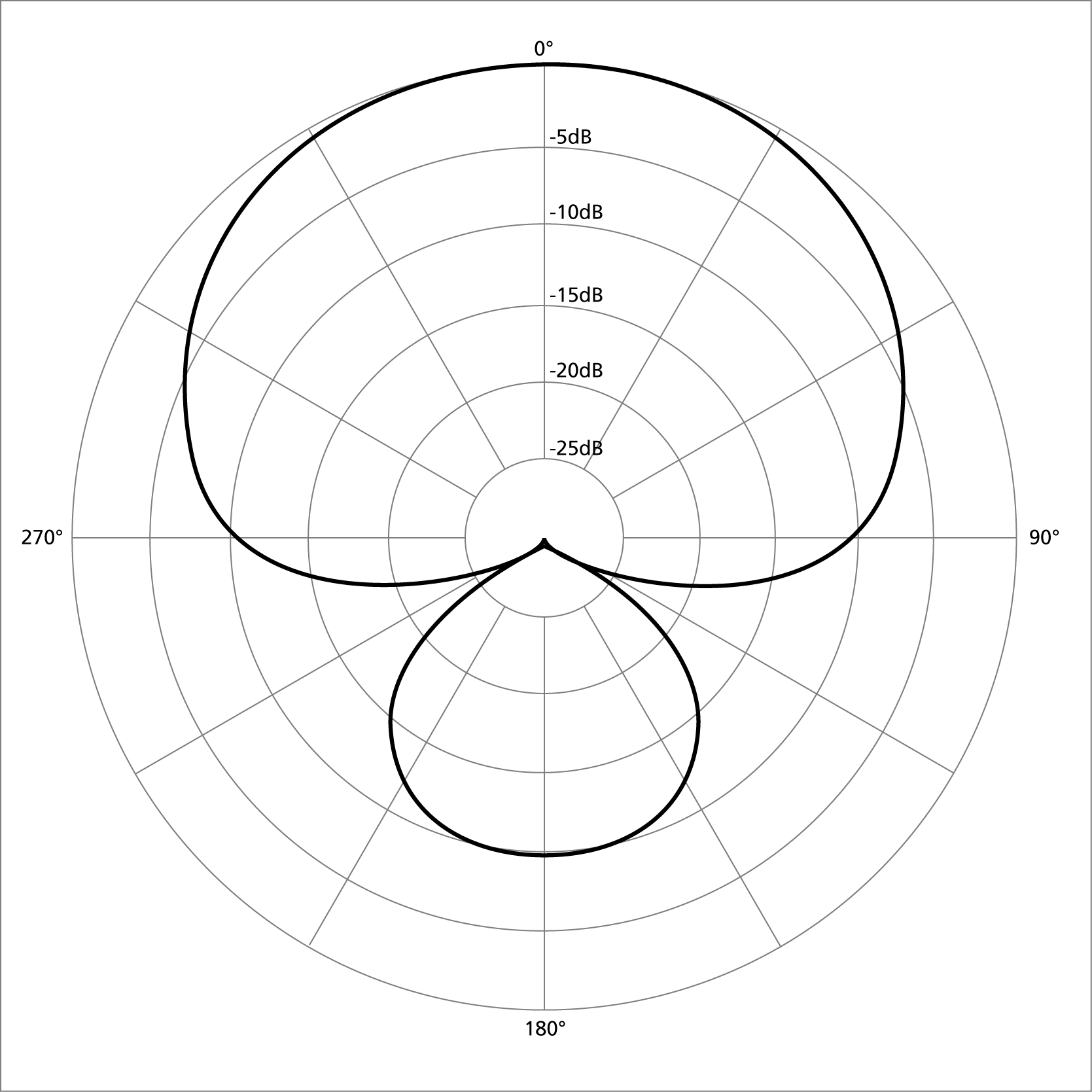
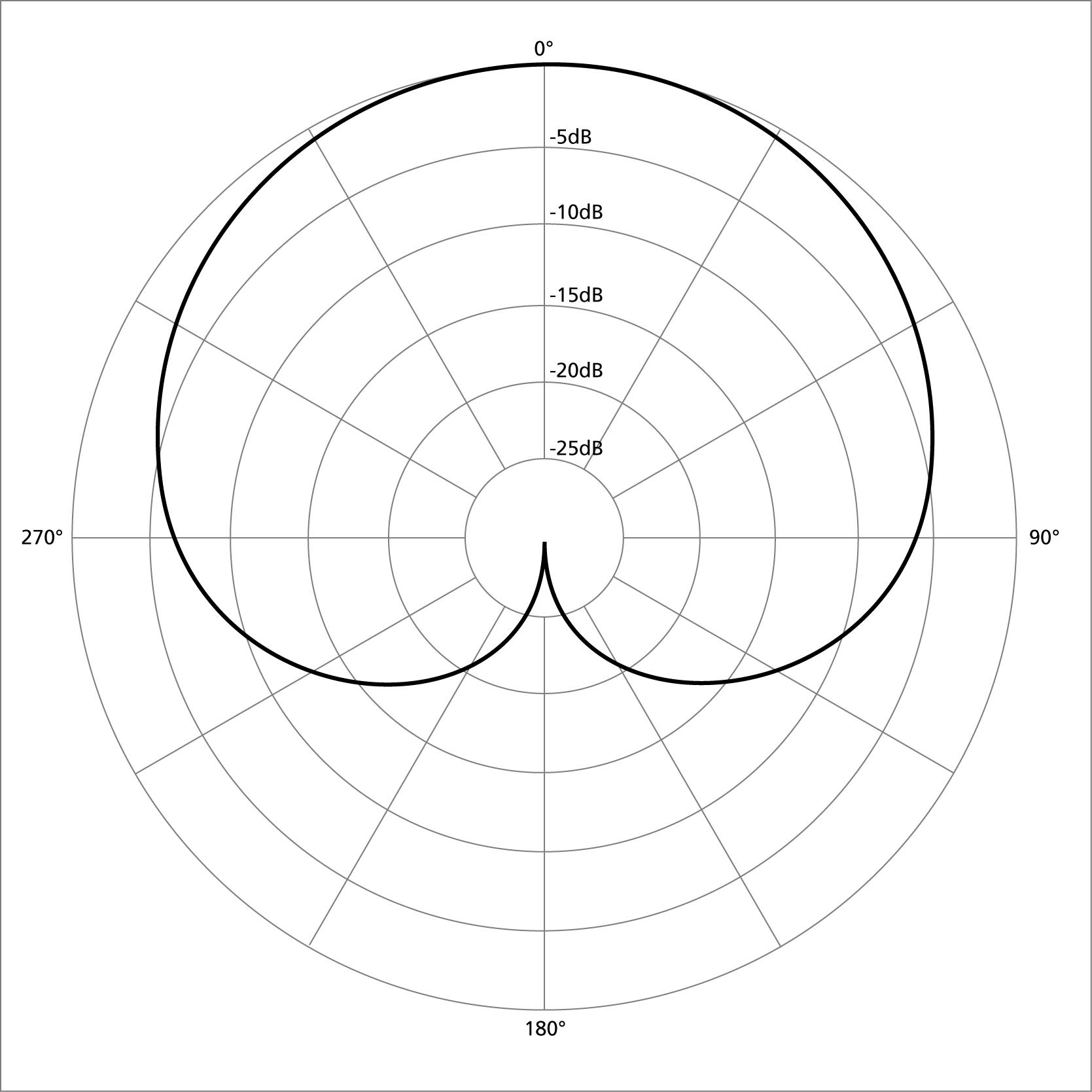
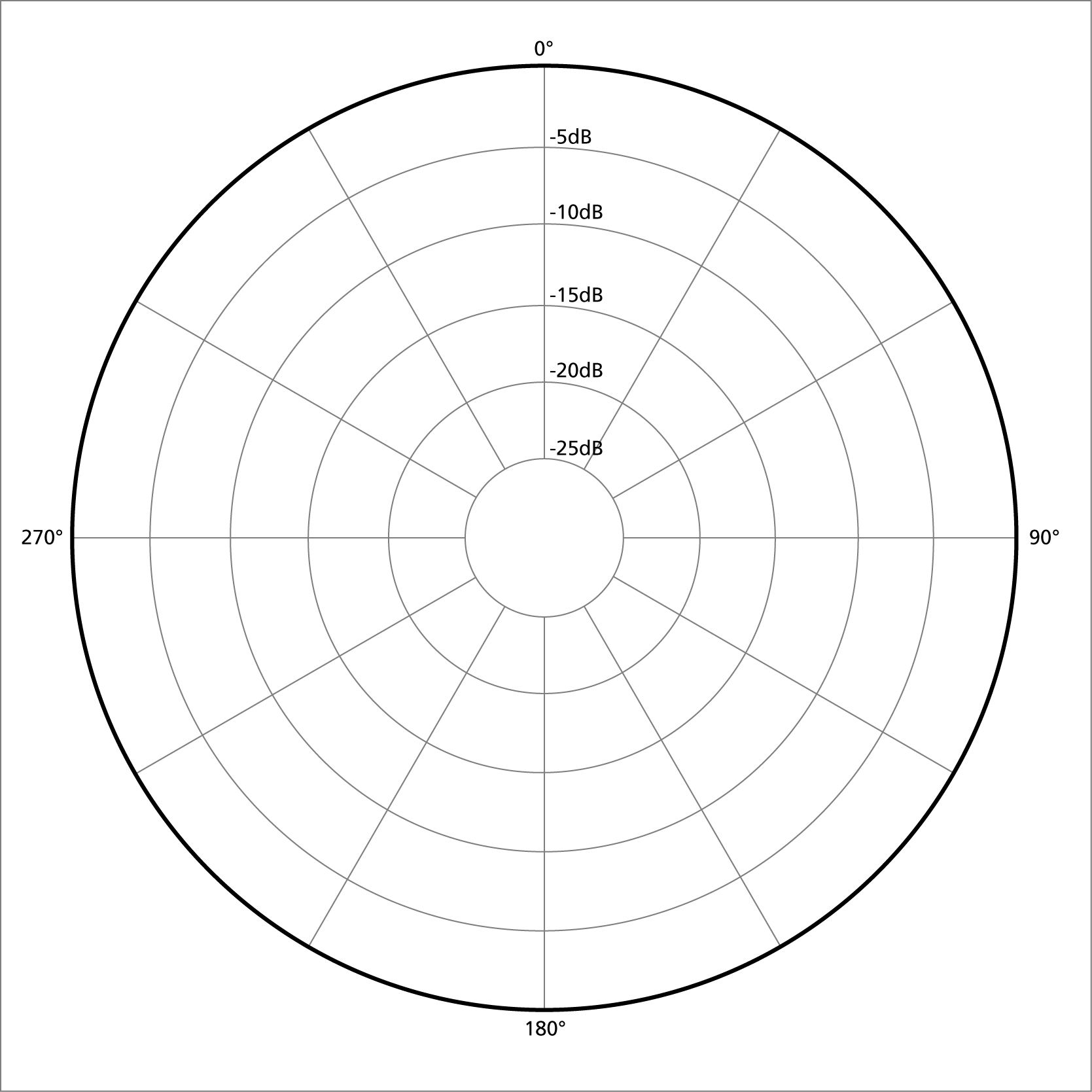


Figure 4: Four common polar patterns of microphones: (top left) omnidirectional, (top right) cardiod, (bottom left) supercardiod, and (bottom right) shotgun (images from Wikimedia Commons).

A microphone’s most important characteristic is its directionality, which indicates how sensitive it is to sounds arriving at different angles around its central axis. A microphone’s directionality is usually represented by a polar pattern showing the location of points that produce the same signal level output in the microphone if a constant sound pressure level is generated from that point.

Figure 4 shows some idealized example patterns. The patterns are considered idealized because in the real world, polar patterns are a function of frequency. Manufacturer diagrams therefore usually include multiple plots at different frequencies. Also, although an omnidirectional microphone's response is generally considered to be a perfect sphere in three dimensions, in the real world this is not the case. The microphone’s body is not infinitely small and, as a consequence, it tends to get “in its own way” with respect to sound arriving from the rear, causing a slight flattening of the polar response. This flattening increases as the microphone’s diameter (assuming it's cylindrical) reaches the wavelength of the frequency in question. Therefore, the smallest-diameter microphone will give the best omnidirectional characteristics, especially at high frequencies.

Different microphone properties result in different applications.

*Headset* and *lavalier* microphones are made for hands-free operation. These are small microphones worn directly on the body. Originally, they were held in place with a lanyard worn around the neck; now, they are typically fastened to clothing with a clip, pin, tape, or magnet. These directed microphones allow mobile use for voice recording. They are in everyday use for videoconferencing, personal recording, theatrical performances, and dictation applications.

A *parabolic* microphone uses a parabolic reflector to collect and focus sound waves onto a microphone receiver, very similar to a satellite dish. Typical uses of this microphone, which has unusually focused front sensitivity and can pick up sounds from many meters away, include nature recording, outdoor sporting event recording, and eavesdropping. Because these microphones tend to have poor low-frequency response as a side effect of their design, they are not typically used for standard recording applications. However, machine intelligence might be able to infer information from them (for example, in connection with a surveillance camera).

*Noise-canceling* microphones have a highly directional design intended for noisy environments when direct attachment to the body is not desirable. For example, a vocalist might use this type of microphone on a loud concert stage **//wording okay?//**. Often, noise-canceling microphones combine signals received from two membranes that are in opposite electrical polarity or are processed electronically later. The main membrane is mounted closest to the intended source and the second is positioned farther away from the source so that it can pick up environmental sounds to be subtracted from the main signal by destructive interference.

Arrays of *omnidirectional* microphones are best to pick up as much sound from the environment as possible. These are typically used for auditory scene analysis, where objects can be located by analyzing the time delay of arrival between microphones (due to the speed of sound). In addition, combining the signals from a larger set of microphones can enhance the signal quality. This technique is often used in speech recognition, when head- or body-mounted microphones are not desirable.

The electric current output by a microphone, and possibly amplified and/or mixed by further equipment, is a continuous electrical signal, with the voltage **//is//** directly proportional to the sound pressure. In practice, sound recording has a linear area for certain sound pressure levels and frequency ranges and nonlinear areas if the sound pressure and/or the captured frequency is too low or too high. If the sound pressure level is too low, the signal will mostly just be zero; if it is too high, it will reach an internal clipping point (which in the worst case is a short circuit) and will be severely distorted. Even if it does not reach the clipping point, the nonlinear behavior of sound processing devices will lead to distortion when the captured signal is outside the nonlinear scope. This is referred to as the signal being *overdriven*. When the signal is outside the recording device’s linear frequency range, harmonic distortion is introduced. For example, a sine curve might be converted into a much less regularly shaped signal. Figure 5 shows an example.

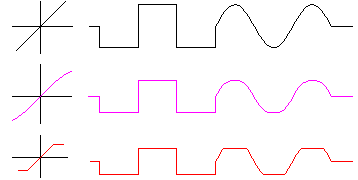


Figure 5. The diagrams on the left show the amplification behavior, the curves on the right show the results for different input signal shapes. Top: Linear behavior, original signal, second row: typical analog behavior and distorted signal due to overdrive, bottom: typical digitization behavior and distorted signal from clipping.

A microphone’s output is usually amplified using an analog amplifier before it is digitized. Some microphones already output a digital signal directly as standardized by the AES 42 **//reference?//** standard.

**//moves in this section okay?//**

**[B Head] Digitization of Sound**

Even if standardized, all sound processing devices have slightly different linear ranges. Sound cables, especially when very long, can also inhibit certain frequencies and, because they often work as “involuntary antennas,” might introduce electric distortion from the outside, the most current one being a “buzz” from the 50 Hz/60 Hz electrical system. Then, recording media, such as old vinyl records or audio cassettes, introduce their own nonlinearities and the effects stack up with each copy made. Therefore, in the last two decades, sound processing has shifted from analog to digital. At the time of this writing, many microphones, mixers, and pre-amplifiers are still analog, but storage and processing is digital. With standards such as AES 42 growing increasingly popular, digitization will soon become a much earlier part of the processing chain.

Digitizing is the representation of a signal by a discrete set of its samples, as Figure 6 shows. Instead of representing the sound signal by an electrical current proportional to its sound pressure, the signal is represented by on-off patterns representing sample values of the analog signal at certain fixed points. The on-off patterns are much less susceptible to distortions than analog signals. Copying in particular is usually lossless. Conceptually, digitization works in two parts, illustrated in Figure 6 **//It’s not really clear from the figure how digitization works in two parts. Perhaps you could add some text to the figure caption to explain what’s happening.//**.

* In *discretization*, **//what?//** reads the analog signal at regular time intervals (the *sampling rate*), obtaining the signal’s value at each interval. Each reading is called a *sample*.
* In *quantization*, samples are rounded to a fixed set of numbers (such as integers).

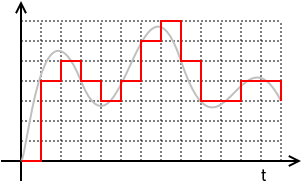


Figure 6. Digital representation of an analog signal. Both amplitude and time axis are discretized.

By generating the signal represented by each sample, we could transform a series of quantized samples back into an analog output that approximates the original analog representation. The sampling rate and the number of bits used to represent the sample values determine how close such an approximation to the analog signal a digitization will be.

The error introduced by the quantization is the *quantization noise*. It affects how accurately the amplitude can be represented. If the samples have very few bits **//correct?//**, the signal will only be represented coarsely, which will both affect the sound’s perceived dynamic and introduce high-frequency artifacts. Typical bit representations for audio are 8, 16, and 24 bits.

The error introduced by the sampling rate is the *discretization error*. This error determines the maximum frequency that can be represented in the signal. This upper frequency limit is determined by the *Nyquist frequency*, named after the Swedish-American engineer Harry Nyquist, or the Nyquist–Shannon sampling theorem, which is half the sampling frequency of a discrete signal processing system. In other words, if a function *x(t)* contains no frequencies higher than *B* hertz, *x(t)* is completely determined by giving its ordinates at a series of points spaced 1/(2B) seconds apart. The proof of this fundamental theorem can be found in the research papers at the end of this chapter.

To illustrate the necessity of *fs* > 2*B*, consider the sinusoid:



With *fs* = 2*B* or equivalently *T* = 1/(2*B*), the samples are given by



These samples cannot be distinguished from the samples of



But, for any *θ* such that sin*(θ) ≠ 0, x(t)* and *y(t)* have different amplitudes and**//a//** different phase, as Figure 7 illustrates.

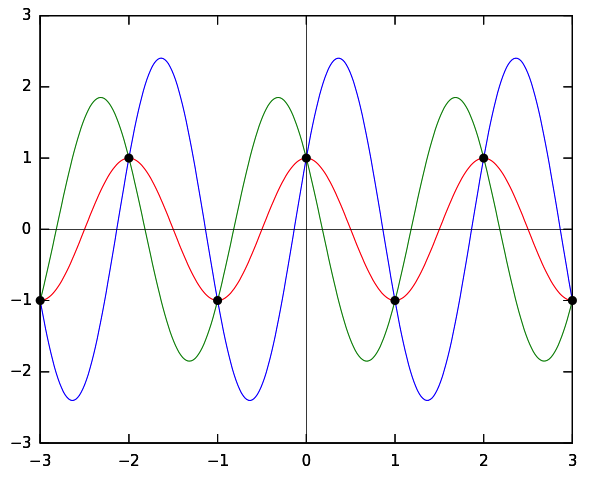


Figure 7. Three possible analog signals for the same sampling points.

The Nyquist **//frequency//** is not limited to sound signals; it is true for the digitization of any signal. However, sound formats are most influenced by this limit. Because the maximum frequency perceptible by the human auditory system is about 22 kHz, compact discs sample at 44 kHz. Human speech, which usually peaks between 6 and 8 kHz, is considered completely represented by a 16-kHz sampling frequency. Professional audio recording equipment frequently use sampling frequencies about 44 kHz, such as 48 kHz and 96 kHz. If the analog equipment supports it, these devices can capture frequencies that are imperceptible by the human ear, allowing for better reproduction of overtones. Further processing, such as digital filters and machine learning, might use the higher frequencies too **//correct?//**.

**[B Head] Reproduction of Sound**

Up to this point, we have assumed the existence of a sound signal. This assumption is generally a safe one, because sound pressure levels can be measured virtually anywhere on earth **//changes okay?//**. However, reproducing sound from storage requires special devices. The most common sound reproduction device is the loudspeaker.

A loudspeaker (or speaker) is the exact reverse of a microphone. A speaker is an electroacoustic transducer that converts an electrical signal into sound. The speaker pulses **//according to?//** the variations of an electrical signal and causes sound waves to propagate through a medium such as air or water. A speaker usually consists of a membrane that is driven back and forth and made to oscillate using an electrical-to-mechanical force converter, such as an electromagnet or a piezo crystal. This core part is typically called the *driver*. The term “loudspeaker” can therefore refer to individual drivers or to an integrated system of drivers in an enclosure. The role of the enclosure, apart from providing a place to mount the drivers, is to prevent sound waves emanating from the back of a driver from interfering destructively (that is, by causing cancellation) with those from the front.

To adequately reproduce a wide range of frequencies, most speakers require a combination of drivers with different properties. Each individual driver is then used to reproduce a different frequency range. Common driver types include subwoofers (very low frequencies, typically below 120 Hz), woofers (low frequencies), mid-range speakers (middle frequencies), tweeters (high frequencies), and sometimes supertweeters, which are optimized for the highest audible frequencies. In systems using multiple drivers, a network of electrical filters, called a *crossover*, separates the incoming signal into different frequency ranges and routes them to the appropriate driver. A speaker system with *n* separate frequency bands is called an *n*-way speaker. Typical home audio devices have a three-way speaker system, consisting of a woofer, a mid-range speaker, and a tweeter.

Like microphones, speakers have directionality—that is, their frequency (re)production properties vary in space. Figure 8 shows the directionality of a typical column-shaper home audio system speaker.

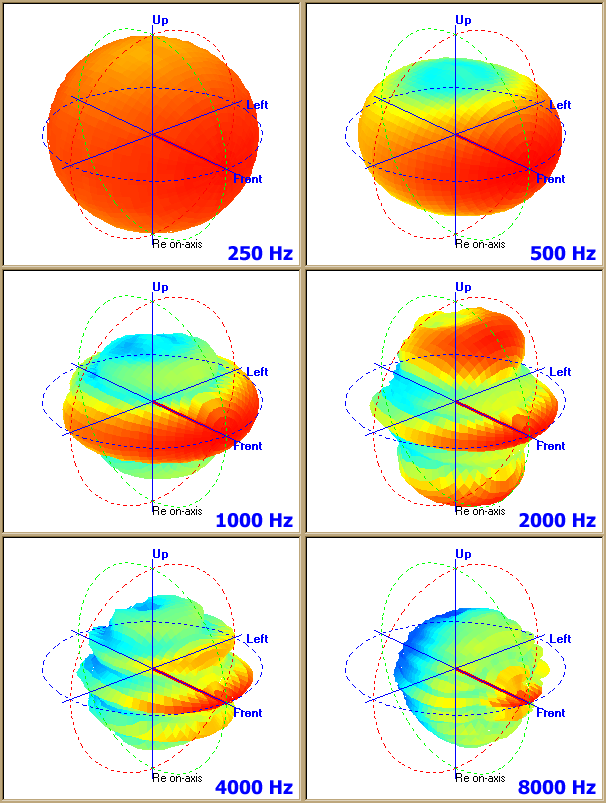


Figure 8. Polar patterns of a typical home speaker system consisting of four drivers at different frequencies. (Source: Wikimedia Commons)

Needless to say, speakers are designed with different directionality for different applications (for example, car speakers have a different polar pattern than supermarket speakers used for making announcements). Other factors that determine a speaker’s properties are

* the rated power, which determines the maximum input a speaker can take before it is destroyed;
* the maximum sound pressure level (SPL), which defines how much sound pressure the speaker can emit;
* the impedance, which determines the electrical compatibility with different amplifiers;
* the crossover frequencies, which define the nominal frequency boundaries of the signal **devision** **//division? decision?//** by the drivers; and
* the frequency range, which determines the speaker system’s linear frequency response range.

The enclosure type (such as sealed or bass reflex) determines some of the loudspeakers’ perceptual properties. Another important factor for sound reproduction quality is the relationship between the number of channels used (for example, two four, or six), how they have been encoded (for example, stereo or surround), and how the speakers are placed in the room when reproducing sound.

Speakers and microphones are the most variable elements in terms of perceived quality. Except for lossy compression, they are responsible for most of the distortion and audible differences in sound systems. Our practical advice to the reader is to use high-quality headphones when experimenting with sound algorithms.

**[A Head] Production of Sound**

**//Add general statement of introduction to this section?//**

**[B Head] Production of Speech**

The production of random noise is relatively easy, but modulating sound in a way suitable for communication requires sophisticated apparatus. Although humans are not the only species to produce sophisticated sounds, they seem to have developed the most sophisticated expressiveness. Speech is the most important means of communication for humans. This section will introduce the basic anatomy and properties of the human speech generation apparatus as well as a few external factors that influence speech production.

Let’s start with some facts. The frequency range of speech is between 80 Hz and about 5 kHz. The pitch of the human voice is between 120 and 160 Hz for adult males and between 220 and 330 Hz for women and children. Vowels can reach frequencies up to about 5 kHz. Sibilants emit the highest frequencies, which can easily reach the nonaudible spectrum (above 20 kHz). The frequency dynamics of speech are relatively high compared to many other sound sources, such as some musical instruments. In general, the volume of the human voice is limited to the sound energy the human body can produce. At 60 centimeters from the mouth, the human voice can typically reach a sound volume of about 60 dBA. A stronger voice can raise the volume by about 6 dB. Yelling measures about 76 dBA (males) and 68 dBA (females).

The articulatory phonetics field, a branch of linguistics, investigates how sounds are produced by the tongue, lips, jaw, and other speech organs. Almost all **//speech?//** organs have additional functions ~~and are not exclusively used for speech production~~. In addition, of course, different **//speech//** organs have more than one function in speech production, and, to make matters even more complex, the same sounds can be produced by different combinations of organs. Figure 9 shows a schematic of the human speech production apparatus.

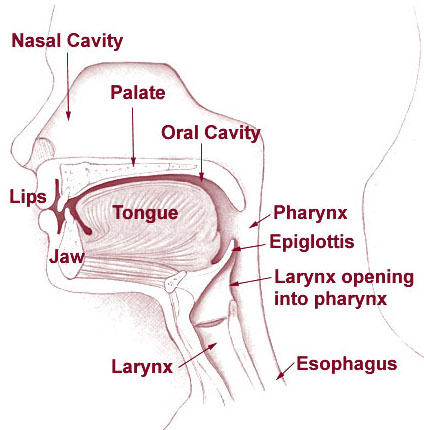


Figure 9. A schematic of the human speech production apparatus. (Source: National Cancer Institute)

Because sound is pressure waves traveling through air, human speech is directly connected to the body’s respiratory system. Speech sounds are usually classified as

* *stop consonants* (with blocked airflow, such as the English pronunciation of “p,” “t,” or “k”),
* *fricative consonants* (with partially blocked and therefore strongly turbulent airflow, such as the English “f” or “v”),
* *approximants* (with only slight turbulence, such as the English “w” and “r”), and
* *vowels* (with full unimpeded airflow, such as the English “a,” “e,” “i,” and “o”).

Other classes exist and there is variation among languages. Vowels are usually classified into *monophtongs*, having a single vowel quality; and *diphtongs*, which manifest a clear change in quality from start to end as in the words bite, bate, or boat.

Consonants and vowels are the building blocks of speech. Linguists refer to these building blocks as *phonemes*. American English has 41 phonemes, although the number varies according to the speaker’s dialect and the system that the linguist doing the classification uses. A phoneme’s concrete pronunciation depends on the previous and the following uttered speech sounds. It also depends on the type of speech (for example, whispering versus screaming) and the speaker’s emotional state, as well as the anatomy of the throat, age, native language and dialect, and social and environmental surroundings. Diseases of the lungs or the vocal cords and articulatory problems, such as stuttering, lisping, and cleft palate, all affect the sound and clarity of speech. In other words, the actual frequency pattern of a specific uttered consonant or vowel has a large variance.

Environmental effects and the brain processing input from other modalities, such as sight or touch, can greatly affect speech. This *Lombard effect* describes an involuntary tendency to increase volume, change pitch, or adjust duration and sound of syllables in response to external noise. This compensation effect increases the auditory signal-to-noise ratio of the speaker’s spoken words. This is one reason why automatic speech recognition algorithms trained in a quiet environment are difficult to transform to noisy environments. For example, an algorithm trained in a developer’s cubicle will rarely work in a car.

The *spectrogram* (see Figure 10) is a standard tool for visualizing and further analyzing sound patterns. A spectrogram is an image that shows how a signal’s spectral density varies with time—that is, it shows the distribution of the energies in different frequency bands in time.

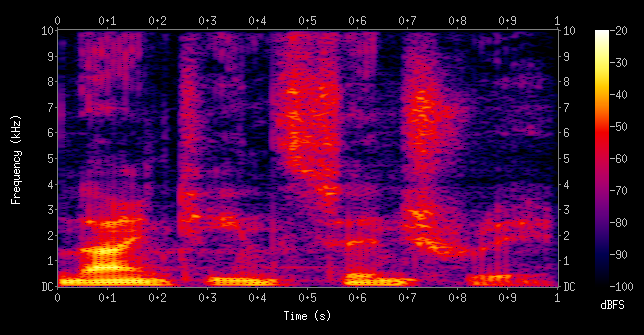


Figure 10. A spectrogram of a male voice saying: “nineteenth century.” The yellow bands of energy are the formants. (Source: Wikimedia commons)

Each phoneme is distinguished by its own unique pattern in the spectrogram. For voiced phonemes, the signature involves large concentrations of energy, the *formants*. Formant values vary widely from person to person. So, reading a spectrogram mainly means recognizing patterns that are independent of particular frequencies and identify the various phonemes with a high degree of reliability.

Relatively static formants are found in the monophthong vowels and the nasals; formants that are more variable over time are found in the diphthong vowels and the approximants. The monophthong vowels, which have the strongest and most stable formants, can usually be easily distinguished by the frequency values of the first two or three formants, which are called F1, F2, and F3. Depending on the phoneme, F1 varies from about 300 to 1,000 Hz, F2 from 850 to 2,500 Hz, and F3 from 2,300 to 3,000 Hz. Higher formants such as F4 and F5 are no longer used for communication, but are indicative of the speaker’s voice. As a result of the low bandwidth and the Nyquist theorem, F4 and F5 are usually lost in telephone speech, as are many of the speakers’ individual voice characteristics.

Unvoiced speech sounds are not usually said to have formants. Still, plosives are usually recognized as a great burst of energy across all frequencies occurring after a short relative silence. Aspirates and fricatives are recognized as “large clouds” of smooth energy along both the time and frequency axes.

**[B Head] Production of Music**

Researchers have discovered archaeological evidence of musical instruments dating as far back as 37,000 years ago. The building and use of musical instruments vary with history and culture as do the sounds that these instruments produce and the musicians playing them. For multimedia computing, we are interested in determining instruments’ general properties so we can leverage them for compressing audio, detecting instruments, and manipulating or artificially synthesizing recordings.

The *fundamental frequency*, abbreviated as f0 or F0 (speak: f-zero), is the inverse of a period length of a periodic signal. Pitch represents a sound’s perceived fundamental frequency. Although the fundamental frequency can be precisely determined through physical measurement, it might differ from the perceived pitch because of overtones. An overtone is either a harmonic or partial (nonharmonic) resonance. In most musical instruments, the frequencies of these tones are close to the harmonics. The harmonic of a wave is a component frequency of the signal that is an integer multiple of the fundamental frequency. For example, when the fundamental frequency is f, the harmonics have frequencies f, 2f, 3f, 4f, and so on. The most important property of the harmonics is that they are all periodic at the fundamental frequency. In other words, the sum of the harmonics is also periodic at that frequency.

*Timbre* describes the quality of sound that distinguishes different types of sound production, such as different musical instruments. The frequency spectrum and time envelope are two physical characteristics of sound that mediate the perception of timbre.

*Spectrum* is the sum of the distinct frequencies emitted by an instrument playing a particular note, with the strongest frequency being the fundamental frequency. When an instrument plays a tuning note (for example A = 440 Hz), the emerging sound is a combination of frequency components, including 440 Hz, 880 Hz, 1,320 Hz, and 1,760 Hz (harmonics), and some partials. The relative amplitudes of these different spectral components is responsible for each instrument’s characteristic sound.

The model typically used to describe a timbre’s time envelope divides sound development into four stages: *attack* (the time from when the sound is activated to its reaching full amplitude), *decay* (the time the sound needs to drop from maximum amplitude to sustain level), *sustain* (the volume level the sound is at until the note is released), and *release* (the time needed for the sound to fade when the note ends). This is also known as the *ADSR envelope.* Psychoacoustics uses the word tone quality and tone color as synonyms for timbre.

The three main categories of musical instruments in the western world are string, wind, and percussion. A string instrument, such as a violin or a guitar, produces sound by vibrating strings. The strings’ vibrations have the form of standing waves that produce a single fundamental frequency (pitch) and all harmonics of that fundamental frequency simultaneously. These frequencies depend on the string’s tension, mass, and length. The harmonics make the sound timbre fuller and richer than the fundamental alone. The particular mix of harmonics present depends on the method of excitation of the string, such as bowing or strumming, as does the timbre. The sound timbre is also significantly affected by resonances in the body of the instrument itself.

A wind instrument contains some type of resonator, usually a tube, in which a column of air is set into vibration by the musician blowing into the end of the resonator. The length of the tube determines the vibration’s pitch. The length is usually varied artificially by manual modifications of the effective length of the vibrating column of air—for example, by covering or uncovering holes in the tube. The sound wave travels down the tube, reflects at one end, and comes back. It then reflects at the other end and starts over again. For a note in the flute’s lowest register, for example, the round trip constitutes one cycle of the vibration. The longer the tube, the longer the time taken for the round trip, and so the lower the frequency.

A percussion instrument produces sound by being hit, shaken, rubbed, scraped, or any other action that sets it into vibration directly. The acoustics of percussion instruments is the most complex because most percussion instruments vibrate in rather complex ways. In general, at low-to-medium amplitudes, their vibrations can be conveniently described by the terms introduced in this chapter. At large amplitude, however, they might show distinctly nonlinear or chaotic behavior. Percussion instruments have the highest variance in frequency and amplitude range and are therefore the most difficult to process.

Many musical pieces contain a mixture of instruments, including human voices. Once mixed, separating the individual instruments would require an adequate model of each instrument’s behavior in its environment and with the recording equipment used. For this reason, music is not only recorded and digitized but also saved in a note-like format, called MIDI, that defines a protocol to control electronic instruments. Electronic instruments have long tried to mimic traditional ones through a process called music synthesis, which we briefly describe next.

**[B Head] Synthesis of Sound**

The artificial generation of speech and music is called *synthesis*. The first music synthesizers date back to 1876. Then, as today, the main goal was not necessarily to correctly imitate a physical musical instrument. Often, the goal was to create new sounds of artistic value. The difficulty and complexity of exact simulation of a real instrument depends of course on that instrument’s properties. It’s easier to simulate a flute than a piano or an organ. It’s not unusual for algorithms to be invented for a particular subtype of instrument. In general though, modern music synthesis is performed by physical modeling of the instrument as well as incorporating original samples of the instrument—the *wavetables*.

Research has recently converged to apply these synthesis techniques to speech **//changes okay?//**. Synthesized speech is often created by concatenating pieces of recorded speech from a database. Systems currently differ in the size of the stored speech units. A system that stores phones or tuples of phones (*diphones*) provides the largest output range but might lack clarity and naturalness in the voice output. Trading off output range for usage in a specific domain the storage of entire words or even sentences allows for higher quality output. The database is usually combined with a model of the vocal tract (such as LPC, see chapter XXX) and other human voice characteristics to create a completely synthetic voice output. This concept of *adaptive concatenative sound synthesis* is the same as for both speech and music synthesis.

**Exercises**

1. How many dBs are 50%, 1%, 0.01%, and 200%? How many dBs can be stored in 16 bits, 24 bits, and 32 bits?
2. List the factors that would influence echo and reverberation in a lecture hall.
3. You are a researcher on a project that requires you to make frequent sound recordings. Unfortunately, your officemate needs to have a very noisy server farm standing beside him. Given no social rules or limitations, what would be the best thing to do to isolate the noise?
4. Discuss what would be the best directionality for a microphone that is used for field studies where you interview people in noisy environments.
5. Discuss and experiment with ideas to reconstruct frequencies beyond the Nyquist limit. What are the trade-offs?
6. Explain how the Nyquist limit sometime becomes visible in the image and video domain. What are the typical artifacts?
7. Explain the artifacts you would expect from a microphone/loudspeaker that is forced to record/play sound both (a) outside its frequency range, and (b) outside its amplitude range.
8. Assume you want to record a classroom seminar with many participants. What environmental noise would you expect?
9. When a signal is received by a microphone, it is amplified and passed out of a loudspeaker. The sound from the loudspeaker might be received by the microphone again, amplified further, and then passed out through the loudspeaker again. The effect is known as *Larsen effect* or, more colloquially, as the *feedback loop*. Describe what happens and what the signal looks like.
10. Which differences would you expect to see in a spectrogram between male and female speakers? Which would you expect to see between younger and older speakers?
11. What is the typical spectrogram of a flute, a violin, or a drum?
12. Implement an ADSR envelop filter and play around with it. Apply it to different sounds and waveforms, including noise.

Literature

*Data Interchange on Read-only 120 mm Optical Data Disks* (CD-ROM). ECMA-130. June 1996.

F. Alton Everst, Ken Pohlmann. *Master Handbook of Acoustics*, McGraw-Hill, 5th edition, June 2009.

D. T. Blackstock. *Fundamentals of Physical Acoustics*, Wiley-Interscience, 1st edition, February 2000.

B. Gold, N. Morgan. *Speech And Audio Signal Processing: Processing And Perception Of Speech And Music*, John Wiley & Sons, 1st edition, August 2006.

A. C. Bickford, R. Floyd. *Articulatory Phonetics: Tools for Analyzing the World’s Languages*, 4th edition, SIL International, July 2006.

T. Kientzle. *A Programmer’s Guide to Sound*, Addison-Wesley, October 1997.

Web Links

MIDI specifications: <http://www.midi.org>/

Neck anatomy: <http://training.seer.cancer.gov/head-neck/anatomy/overview.html>

Research Papers

H. Nyquist. ”Certain topics in telegraph transmission theory,” *Trans. AIEE*, vol. 47, pp. 617-644, Apr. 1928. Reprint available as classic paper in: *Proceedings of the IEEE*, Vol. 90, No. 2, Feb 2002.

É. Lombard. “Le signe de l'élévation de la voix,” *Annales des Maladies de L’Oreille et du Larynx*, Vol. XXXVII, No. 2, pp. 101–119, 1911.

Slabbekoorn H, Peet M. “Birds sing at a higher pitch in urban noise.” *Nature*, 424(6946):267, 2003.

Junqua JC. “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *Journal of the Acoustic Society of America*, Jan;93(1):510-24, 1993.

Scheifele PM, Andrew S, Cooper RA, Darre M, Musiek FE, Max L. “St. Lawrence River beluga Indication of a Lombard vocal response in the St. Lawrence River Beluga,” *Journal of the Acoustic Society of America*, 117(3 Pt 1):1486-92, 2005.